

# How to merge three different methods for information filtering ?

## Abstract

Twitter is now a gold marketing tool for entities concerned with online reputation. To automatically monitor online reputation of entities, systems have to deal with ambiguous entity names, polarity detection and topic detection. We propose three approaches to tackle the first issue: monitoring Twitter in order to find relevant tweets about a given entity. Evaluated within the framework of the *RepLab-2013 Filtering task*, each of them has been shown competitive with state-of-the-art approaches. Mainly we investigate on how much merging strategies may impact performances on a filtering task according to the evaluation measure.

## 1 Introduction

Online reputation is a key information for public figures and companies in order to react to the public opinions and to anticipate them. Indeed, knowing what make their reputation good or bad allows them to make informed decision. For instance a company may make additional efforts on its call centers if it notices that its consumers are unsatisfied.

Monitoring online reputation of entities requires to be able to retrieve all opinions or reviews about them. Automatic approaches have then to deal with the noise generated by the recall-oriented retrieving techniques used. This noise is mainly the result of entity names ambiguity (e.g. *jaguar* which may refer to an animal or a car manufacturer). A classification step is required to filter out sources which do not actually mention the monitored entity. Topic

detection is necessary to identify which matter is discussed in the source and finally the polarity of it has to be estimated (is the opinion positive, neutral or negative?). Each of these three issues is an open problem. Moreover, systems have to be able to process large amounts of incoming new documents in a short time to provide fresh feedbacks. Sources commonly used are news web sites, blogs, forums or more recently social networks such as Twitter.

We propose three approaches to filter tweets on whether or not they refer to a given entity. These approaches rely on tweets content and meta-data associated to them (timestamp, user names, ...) as well as on the information contained about the entity in a knowledge base. We also investigate if combining systems outputs with merging algorithms can improve the overall performances and if different strategies may be applied to promote a measure or another.

Each proposition is evaluated within the framework provided by the RepLab 2013 evaluation campaign and they all show competitive performances.

The remainder of this paper is organized as follows. Section 2 presents related works. Section 3 describes the proposed systems. Section 4 gives details about merging algorithms we used. Experiments are described in Section 5. In Section 6 we discuss the results before concluding in section 7.

## 2 Related Work

A decade ago, a TREC task called "Filtering" (Robertson and Soboroff, 2002) had the following definition: finding documents relevant to a query in a stream of data. Effective approaches

were inspired by information retrieval techniques to score documents (Okapi (Robertson and al, 2002), Rocchio (Schapire and al, 1998), ...).

In 2012, a new TREC task called "Knowledge Base Acceleration" (KBA) (Frank and al, 2012) started with a more entity centric definition: filtering a time-ordered corpus for documents that are related to a set of entities from Wikipedia. The best performing approach used one classifier (SVM) by entity tracked with features representing whether or not a term is in a document, regardless of its frequency (Kjersten and McNamee, 2012). Training data have however to be provided for each new entity "followed". Another successful approach capture intrinsic characteristics of related documents by relying on document centric features, entity profile related features and time features (Bonney and al, 2013).

Recently information filtering on Twitter emerged. (Lee, 2012) for instance followed the evolution of big and short terms events, like natural disasters, in real-time. RepLab 2012 Filtering task (Amigò and al, 2012) follows the KBA 2012 definition but focus on Twitter as the source of incoming data (instead of news, blogs and forum posts). The submitted approaches rely on various sources of evidence like named entity recognition (Villena-Román and al, 2012), matches of terms between tweets and Wikipedia (Younus and al, 2012) or the importance of features specific to Twitter such as the presence of a user name in a tweet (Peetz and al, 2012) or the number of hashtags (Chenlo and al, 2012).

Merging metrics or methods used in natural language processing (NLP) and information retrieval can be seen, as shown in (Lamontagne and Abi-zeid, 2006), as a multi-criteria optimization problem: in particular, the ELECTRE methods (Figueira and al, 2005), which turned out to be efficient applied to industrial domains (Gourion and Josselin, 2012), have been transposed to an NLP context (Carrillo and al, 2012) opted for a voting method to combine their runs with (Chenlo and al, 2012).

## 3 Methods

### 3.1 Cosine distance (TF-IDF-Gini)

The first approach consists in a supervised classification based on a cosine similarity. Vectors used to compute similarities are built using the Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988) and the Gini purity criterion (Torres and al, 2012).

Tweets are cleansed by removing hypertext links and punctuation marks, hashtags and @ before a user name. We have removed a set of tool-words and some entities ID. Terms are lower-cased. We generate a list of n-grams by using the Gini purity criterion.

We create terms (words or n-grams) models for both classes (related and unrelated tweets) and term frequencies are computed with the TF-IDF and Gini criterion. These models take into account the following meta-data: user id, entity id and language integrated as terms in the bag-of-terms of tweets.

A cosine similarity measures the distance between the bag-of-terms of a tweet and the whole bag built for each class and ranks tweets according to this measure.

### 3.2 KNN with discriminant features

The system tries to match each tweet in the test set with the  $K$  most similar tweets in the training set. Tweet similarity is computed using Jaccard measure on the bag-of-words discriminant representation of the tweets. As in section 3.1, each tweet is represented as a vector whose components are weighted according to TF-IDF and the Gini purity criterion. The process also takes into account tokens created from the meta-data (author, entity-id). The stoplist of section 3.1 has been used.

### 3.3 Adaptation KBA'12 system

For the KBA filtering task, a state-of-the-art approach consist in capturing intrinsic characteristics of highly relevant documents by mean of three types of features: document centric features, entity's profile features, and time features (Bonney and al, 2013). Features are computed for each candidate document and, using a Random Forest classifier, used to determine if the document is related or not to a given entity.

Unlike previous approaches it doesn't require a new set of examples for each new entity. We want to measure the robustness of this approach by using it on another type of documents (i.e. tweets). No adjustments are made on it but tweets are however pre-processed: stop-words are deleted as well as @ before user names and hashtags are split. The classifier is trained on all related and unrelated examples for each type of entities (automotive, universities, banking and music/artists).

## 4 Merging algorithms

To improve the performances we use three ways of combining our systems outputs.

### 4.1 Linear combination of outputs score

$N$  systems are available. For each tweet  $T$  of the test set, one system  $j$  associates each label  $L_k$  with a confidence score  $s_j(T, L_k)$  ( $j = 1, \dots, N$ ). The output entity label  $L$  is chosen according to the following rule :

$$L = \arg \max_k \left( \sum_{j=1}^N s_j(T, L_k) \right) \quad (1)$$

### 4.2 ELECTRE I method

The goal of this method (Roy, 1991) is to choose the best label from the entire set of labels ranked according to the different systems.

A relation  $\mathcal{S} \subset \mathbb{L} \times \mathbb{L}$ , denoted "over-ranking", is defined on the label set  $\mathbb{L}$ : a label  $l$  over-ranks another label  $l'$  if  $l$  dominates  $l'$  on an "important" number of systems and if  $l'$  does not dominate "too much"  $l$  on the remaining systems.

More precisely, for each pair of labels  $(l, l')$ , a concordance index  $c(l, l')$  is computed, corresponding to the proportion of systems where  $l$  dominates  $l'$ .  $l$  over-ranks  $l'$  if  $c(l, l')$  exceeds a concordance threshold, generally fixed around  $2/3$  and if  $l$  is not dominated by  $l'$  on the remaining systems above a *veto* threshold, which has been fixed here to  $v = 0.5$ .

The set of the best labels, possibly empty and denoted as the kernel of the relation  $\mathcal{S}$ , consists in the labels which are not overranked by others. If there is no, or more than one, label in the kernel, this method is discarded and the merging algorithm described in

the previous subsection, based on a linear combination of the scores, is applied.

### 4.3 PROMETHEE mono-criterion method

This method relies on a concordance matrix: for each pair of labels  $(l_i, l_j)$ , the matrix coefficient  $c_{ij}$  corresponds to the concordance index  $c(l_i, l_j)$  introduced in the previous subsection.

For each label  $l_i$ , two sums are computed:  $s_l(l_i) = \sum_j c_{ij}$  and  $s_c(l_i) = \sum_j c_{ji}$ .  $s_l(l_i)$  measures the tendency of  $l_i$  to dominate the other labels, and  $s_c(l_i)$  the tendency of  $l_i$  to be dominated.

The final score of the label  $l_i$  is the difference  $s_l(l_i) - s_c(l_i)$  and the dominant label is the one whose score is maximal.

## 5 Experiments

### 5.1 Replab 2013 Framework

The corpus is a bilingual (English and Spanish) collection of tweets containing the name of one of the 61 entities selected in four domains: automotive, banking, universities and music/artists. Tweets have been collected by querying the Twitter search engine<sup>1</sup>. The dataset covers a period going from the 1<sup>st</sup> of June 2012 to the 31<sup>st</sup> of December 2012. 42,700 tweets have been provided for training purpose and 100,000 tweets for the evaluation. The training set is composed of the 700 first tweets retrieved for each entity. For each entity, at least 2,200 tweets have been collected.

Tweets, however, are not homogeneously distributed across the entities.

Systems are evaluated according to the following measures: Accuracy, Reliability and Sensitivity (Amigò and al, 2013). Reliability is defined as precision of binary relationships predicted by the system with respect to those that derive from the gold standard; and Sensitivity is similarly defined as recall of relationships. A F-measure is then used to combine both scores.

These measures are well adapted to the task but are really severe on unbalanced datasets.

### 5.2 Results

Table 1 shows results of our approaches against the official RepLab 2013 baseline and the median sys-

<sup>1</sup><http://twitter.com/search>

| Approach             | Accuracy    | Reliability | Sensitivity | F-Measure   |
|----------------------|-------------|-------------|-------------|-------------|
| MPMS                 | .899        | .668        | <b>.367</b> | <b>.400</b> |
| OTB                  | .902        | .651        | <b>.367</b> | .386        |
| Naive LC             | <b>.904</b> | <b>.691</b> | .364        | .385        |
| Naive Elec           | .903        | .671        | .363        | .383        |
| k-NN                 | .890        | .658        | .357        | .381        |
| KBA                  | .878        | .619        | .331        | .341        |
| R-LC                 | .895        | .680        | .290        | .313        |
| <i>Baseline</i>      | .876        | .461        | .325        | .312        |
| R-Elec               | .892        | .680        | .281        | .302        |
| Cosine               | .834        | .423        | .331        | .272        |
| <i>RepLab Median</i> | .826        | .489        | .286        | ..265       |

Table 1: Results on the Filtering Task ordered according to the F-Measure.

tem among participants.

The baseline<sup>2</sup> is a supervised system that matches each tweet in the test set with the most similar tweet in the training set, and assumes that the annotations in the tweet from the training set are also valid for the tweet in the test set. Tweet similarity is computed using Jaccard distance and a straightforward bag-of-words representation of the tweets.

The method described in section 3.2 can be considered as an improved version of the baseline.

Two systems (KNN and KBA with a F-measure scores of respectively .381 and .341) have reached greater performances than the baseline on every measures. The confidence interval (.002 and .005 respectively for accuracy and F-Measure) computed following *Polling Method* (Voorhees, 1998) shows that the difference between the systems is significant.

Merging strategies R-Elec (for ELECTRE) and R-LC (for Linear Combination) did not produce good selection rules since their performances remain lower than the best system taken alone. A natural merging strategy consisting in merging only the best systems on a development set gives better results (Naive LC and Naive Elec).

Moreover, a multi pass strategy (MPMS) merging systems in pair before considering merging all pairs improves Sensitivity and thus the F-Measure (.400) despite of a loss in term of accuracy and reliability. Finally, merging only the best (OTB) runs on each measure gives quite similar improvements.

These results show that using merging strategies to combine different systems lead to improvements, whatever the metric chosen. The key observation is that it is possible to pick a merging strategy according to the metrics we choose to focus on. A quite naive merging strategy (Naive) seems to result in a better precision (improvements in both Accuracy and Reliability). On the contrary, adopting a multi pass strategy (MPMS) allow to give a highest priority to recall in both classes (i.e. Sensitivity). Finally, if a compromise is preferred, we saw that promoting systems that did well on each measure (OTB) is a good option.

## 6 Conclusion

In this paper we presented some of the interesting features of the systems that we evaluated within the framework provided by RepLab 2013 as well as their performances. We proposed several combinations of them using different merging strategies in order to take benefit from the diversity of information offered by our systems. We also showed that these merging strategies have to be applied depending on the evaluation measures to offer in one hand the best results according to a specific measure or in the other hand to obtain a trade-off. Since a merging strategy cannot get the best score according to each metrics, we can accept a loss according to one metric if it has a real impact on the task official measures.

A more advanced view would be to apply a specific merger entity by entity, especially for unbalanced entities.

<sup>2</sup><http://www.limosine-project.eu/events/replab2013>

## References

- [Davis and al2012] A. Davis, A. Veloso, A. da Silva, W. M. Jr. and A. Laender. Named entity disambiguation in streaming data. *Proceedings of the 50th meeting of the ACL*, 2012.
- [Younus and al2012] A. Younus and C. O’Riordan and G. Pasi. *CIRDISCO at RepLab2012 Filtering Task: A Two-Pass Approach for Company Name Disambiguation in Tweets*. in proceedings of CLEF 2012.
- [Kjersten and McNamee2012] B. Kjersten and P. McNamee. The hltcoe approach to the trec 2012 kba track. *Proceedings of The 21th TREC*, 2012.
- [Roy1991] B. Roy. 1991. *The outranking approach and the foundations of ELECTRE methods..* Theory and Decision, 31 :49–73.
- [Gourion and Josselin2012] D. Gourion and D. Josselin. *Aide à la décision robuste pour la localisation d’un centre de traitement des déchets. Comparaison de méthodes d’analyse multicritères*. Annales de l’ISUP 56 (2012), no 2-3, 17–35.
- [Amigò and al2012] E. Amigò and A. Corujo and J. Gonzalo and E. Meij and M. de Rijke. *Overview of RepLab 2012: Evaluating Online Reputation Management Systems*. in proceedings of CLEF 2012.
- [Amigò and al2013] E. Amigò and J. Gonzalo and F. Verdejo. *A General Evaluation Measure for Document Organization Tasks*. in SIGIR 2013.
- [Voorhees1998] Ellen M. Voorhees. 1998. *Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness*. Information Processing and Management, 36(5), 697–716. 2000.
- [Salton and Buckley1988] G. Salton and C. Buckley. 1988. *Term weighting approaches in automatic text retrieval*, pp 513–523. in Information Processing and Management 24.
- [Carrillo and al2012] J. Carrillo de Albornoz and I. Chugur and E. Amigò. *Using an Emotion-based Model and Sentiment Analysis Techniques to Classify Polarity for Reputation*. in proceedings of CLEF 2012.
- [Chenlo and al2012] J. Chenlo and J. Atserias and C. Rodriguez and R. Blanco. *FBM-Yahoo! at RepLab 2012*. in proceedings of CLEF 2012.
- [Frank and al2012] J. Frank, M. Kleiman-Weiner, D. Roberts, F. Niu, C. Zhang, and C. Ré. Building an entity-centric stream filtering test collection for trec 2012. *Proceedings of The 21th TREC*, 2012.
- [Figueira and al2005] J. Figueira and S. Greco and M. Ehrgott. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer Verlag 2005.
- [Lee2012] J. Lee. Mining spatio-temporal information on micro-blogging streams using a density-based online clustering method. *Expert Systems with Applications*, 2012.
- [Torres and al2012] J M. Torres-Moreno and M. El-Beze and P. Bellot and F. Bechet. 2012. *Opinion detection as a topic classification problem*. in Textual Information Access chapter 9, ISTE Ltd John Wiley and Son.
- [Villena-Román and al2012] J. Villena-Román and S. Lana-Serrano and C. Moreno-García and J. García-Morera and J. Carlos González-Cristóbal. *DAEDALUS at RepLab 2012: Polarity Classification and Filtering on Twitter Data*. in proceedings of CLEF 2012.
- [Bonnefoy and al2013] L. Bonnefoy and V. Bouvier and P. Bellot. *A Weakly-Supervised Detection of Entity Central Documents in a Stream*. in SIGIR 2013.
- [Lamontagne and Abi-zeid2006] L. Lamontagne and I. Abi-zeid. *Combining Multiple Similarity Metrics Using a Multicriteria Approach*. in Proceedings of EC-CBR 2006, LNAI 4106.
- [Peetz and al2012] M. Hendrike Peetz and M. de Rijke and A. Schuth. *From Sentiment to Reputation ILPS at RepLab 2012*. in proceedings of CLEF 2012.
- [Cancedda and al2002] N. Cancedda and C. Goutte and J. M. Renders and N. Cesa-Bianchi and A. Conconi and Y. Li and J. Shawe-Taylor and A. Vinokourov and T. Graepel and C. Gentile. Kernel methods for document filtering. *Proceedings of The 11th TREC*, 2002.
- [Robertson and Soboroff2002] S. Robertson and I. Soboroff. The trec 2002 filtering track report. *Proceedings of The 11th TREC*, 2002.
- [Robertson and al2002] S. Robertson, S. Walker, H. Zaragoza, and R. Herbrich. Microsoft cambridge at trec 2002: Filtering track. *Proceedings of The 11th TREC*, 2002.
- [Schapire and al1998] R. Schapire, Y. Singer, and A. Singhal. Boosting and rocchio applied to text filtering. *Proceedings of the 21st annual international ACM SIGIR*, 1998.
- [Zhang and Callan2001] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. *Proceedings of the 24th annual international ACM SIGIR*, 2001.